Wu and Zhang *BMC Bioinformatics* 2011, **12**(Suppl 9):S7
http://www.biomedcentral.com/1471-2105/12/S9/S7

BMC Bioinformatics

PROCEEDINGS Open Access

Structural properties of the reconciliation space and their applications in enumerating nearly-optimal reconciliations between a gene tree and a species tree

Taoyang Wu*, Louxin Zhang*

From Ninth Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics
Galway, Ireland. 8-10 October 2011

Abstract

Introduction: A gene tree for a gene family is often discordant with the containing species tree because of its complex evolutionary course during which gene duplication, gene loss and incomplete lineage sorting events might occur. Hence, it is of great challenge to infer the containing species tree from a set of gene trees. One common approach to this inference problem is through gene tree and species tree reconciliation.

Results: In this paper, we generalize the traditional least common ancestor (LCA) reconciliation to define a reconciliation between a gene tree and species tree under the tree homomorphism framework. We then study the structural properties of the space of all reconciliations between a gene tree and a species tree in terms of the gene duplication, gene loss or deep coalescence costs. As application, we show that the LCA reconciliation is the unique one that has the minimum deep coalescence cost, provide a novel characterization of the reconciliations with the optimal duplication cost, and present efficient algorithms for enumerating (nearly-)optimal reconciliations with respect to each cost.

Conclusions: This work provides a new graph-theoretic framework for studying gene tree and species tree reconciliations.

Background

With much higher speed than the traditional Sanger sequencing technology, the ultra-deep sequencing technology has made huge amounts of molecular data available for genomics study [1]. It provides an unprecedented opportunity to infer phylogenetic trees from multilocus and genomics data. One approach to inferring phylogeny from multilocus data is to reconstruct a gene tree from each locus and then to combine the resulting trees into a phylogeny, called the containing species tree. Gene trees are often different since each gene family might undergo different mutational events such as gene duplication and loss, horizontal gene transfer, and incomplete lineage sorting [2,3]. Therefore, the containing species tree is inferred from gene trees by reconciling it with each gene tree to minimize the total number of hypothetical evolutionary events that are responsible for the discordance between the trees.

The gene tree and species tree reconciliation was first introduced by Goodman et al. [4] and formally defined by Page [5]. Given a gene tree for a gene family and a containing species tree, a reconciliation between them represents an evolutionary scenario of the gene family within the evolutionary history represented by the species tree [4]. To study gene duplication history, gene

* Correspondence: taoyang.wu@gmail.com; matzlx@nus.edu.sg
Department of Mathematics, National University of Singapore, Singapore 119076
Full list of author information is available at the end of the article

tree and species tree are reconciled to minimize the number of gene duplications and/or losses. The mathematical and algorithmic issues of gene tree and species tree reconciliations have been intensively studied in the past decade [6-14]. For example, it has been shown that the so-called least common ancestor (LCA) reconciliation has the minimum duplication and loss cost [9,15].

Although the LCA reconciliation is optimal in terms of the duplication cost, it may not represent the true evolution of the gene family being considered. Indeed, recent studies suggest that more than one reconciliations may occur with the highest probability [16,17]. Such studies [3,14,17,18] in the stochastic framework assume that the discordance between a gene tree and a species tree is caused by incomplete lineage sorting and adopt Kingman's coalescent theory from population genetics [19].

The fact that the LCA reconciliation may not be the unique optimal with respect to the duplication cost motivates researchers to study the space of all the reconciliations and develop algorithms to enumerate nearly-optimal reconciliations for a species tree and a gene tree [20,21]. In this paper, we take a different approach to these two issues. We generalize the LCA reconciliation to define an arbitrary reconciliation as a vertex-mapping from a gene tree to a species tree that preserves the hierarchical structure of the gene tree. Our approach is essentially different from the existing ones [20,22], where the specific mutation events are used and a gene tree vertex is mapped to a species tree branch to specify a duplication event. One advantage of our approach over the others is that we separate reconciliation concept from the cost models that are used to measure the tree discordance. Because of this, we are able to study the structural properties of the space of all reconciliations between a gene tree and a species tree in the same manner for each of the three cost models. We show that the LCA reconciliation has not only the minimum duplication and loss cost [9,15], but also the minimum deep coalescence cost. We also present a novel characterization of the reconciliations with the optimal duplication cost, and develop efficient algorithms for enumerating (nearly-) optimal reconciliations with respect to each cost model.

Methods

Basic notations

Species evolve from their common ancestor through a series of speciation events. A species tree represents the evolutionary history of a set of species. A gene family might evolve from its common ancestral gene through gene duplication and loss events. Here we will assume that no lateral gene transfer has occurred.

Both gene and species trees are rooted trees with labeled leaves. In a species tree, a leaf x represents a species, the label of x. Hence, the species tree is uniquely leaf-labeled. In a gene tree, a leaf y represents a gene found in a species. To infer the duplication history of a gene family, its gene tree and the containing species tree is reconciled [4]. For this purpose, a leaf of a gene tree is labeled with the containing species. Since a species may contain duplicate genes, two leaves in a gene tree can have the same label.

Let T be a species or gene tree; its vertex set and edge set are denoted by $V(T)$ and $E(T)$, respectively. Given two vertices u and v in T, there exists a unique path $P(u, v)$ from u to v. The number of edges in $P(u, v)$, denoted by $d(u, v)$, is called the *distance* between u and v. Note that $d(u, v) = 0$ if and only if $u = v$. The node v is a *descendant* of u or u is an *ancestor* of v, denoted by $v \le u$, if u is on the unique path from $r(T)$, the root of T, to v. For simplicity, we also write $v < u$ if $v \le u$ and $v \ne u$. Given a set A of vertices in T, u is a *common ancestor* of A if and only if $v \le u$ for every $v \in A$. In addition, if $u \le u'$ for any other common ancestor u' of A, then we say u is the *least common ancestor* of A, written as lca(A), or lca(u_1, ..., u_k) if $A = \{u_1, ..., u_k\}$.

For each vertex u in T with $u \ne r(T)$, the *parent* of u, denoted by $p(u)$, is the unique vertex in T that is adjacent to u and contained on the path from $r(T)$ to u. In this case, u is also called a *child* of $p(u)$. The out-degree of u, denoted by $d(u)$, is defined as the number of the children of u. Obviously, a node is a leaf if and only if its out-degree is 0. Non-leaf nodes are internal nodes; they form a subset $V^o(T)$ of $V(T)$. If every internal vertex has out-degree two, then T is *binary*. For an internal vertex u in a binary tree, its two children are denoted by u_1 and u_2, unless stated otherwise. In this study, we will focus on the case that gene trees and species trees are binary. For a vertex u, we use $L(u)$ to denote the set of the labels of its leaf descendants and call it the *cluster* induced by u. Finally, we use $L(T)$ to denote the set of leaf labels, i.e., the cluster induced by the root of T.

Reconciliation between gene tree and species tree

Let S be a species tree over a set of species and G a gene tree such that $L(G) \subseteq L(S)$, i.e., G is over all the homologous genes of a gene family found in some species. A map f from $V(G)$ to $V(S)$ is order-preserving if for each pair of vertices u, v in G, $u \le v$ implies $f(u) \le f(v)$; it is leaf-preserving if, for each leaf x in G, $f(x)$ is the unique leaf in S that has the same label.

A *reconciliation* between a gene tree G and a species tree S is a leaf-preserving and order-preserving map from $V(G)$ to $V(S)$. Clearly, a reconciliation f between G and S is necessarily an inclusion-preserving mapping (see [8]), that is, for each pair of vertices u, v in G, $u \le$

v implies $L(f(u)) \subseteq L(f(v))$. However, the reverse statement is not true. For instance, the mapping that maps each vertex of G to the root of S is an inclusion-preserving mapping, but according to our definition, it is not leaf-preserving, and hence not a reconciliation.

Note that our definition is consistent with the one used in [20], where a reconciliation is defined as a mapping from $V(G)$ to $V(S) \cup E(S)$ that satisfies three constraints: base constraint, tree mapping constraint and ancestor consistency constraint. Roughly speaking, our order-preserving condition corresponds to the ancestor consistency constraint, and the leaf-preserving condition is related to the base constraint, while the tree mapaaping constraint is not needed in our setting. The main difference between these two frameworks is the model used to interpret mappings. For example, in [20], a duplication event is associated to a vertex v in G if and only if v is mapped to an edge, while in our model, whether v is associated with a duplication event is not solely determined by the image of v.

A reconciliation represents a hypothetical evolutionary history of the gene family. In a gene tree, an internal vertex u represents the common ancestor of the genes represented by the leaves below it. The property just reflects the intuitive fact that u is an ancient gene appearing in some common ancestor of the species from which the genes are taken. Recall that in species tree each branch represents an ancestral species. Under the reconciliation f, we considered u as the gene ancestor found in the species represented by the branch entering $f(u)$.

There is a canonical partial order $\preccurlyeq$ on the set of reconciliations between G and S: for any f' and f, $f' \preccurlyeq f$ if and only if $f'(v) \preceq f(v)$ holds for every vertex v in G. Define a mapping M from G to S recursively as:

$$M(u) = \begin{cases} \text{the unique leaf with the same label,} & \text{if } u \text{ is a leaf;} \\ \text{lca}(M(u_1), M(u_2)), & \text{otherwise.} \end{cases}$$

M is called the *least common ancestor* (LCA) reconciliation between G and S. Note that we have $M \preccurlyeq f$ for every reconciliation f between G and S, because it is easy to see that $M(u) \preceq f(u)$ holds for all $u \in V(G)$, by a bottom-up traversal.

Inference of gene duplications

If the discord of a gene tree G and its containing species tree S is due to gene duplication, a reconciliation f between them represents a plausible duplication history of genes. For an internal vertex u, a *duplication event* is associated with u if and only if one of the following two conditions holds: (D-i) $f(u) = f(u_1), f(u) = f(u_2)$ or both hold; (D-ii) $P(f(u), f(u_1))$ and $P(f(u), f(u_2))$ contain a common edge. In the literature (see [5]), when the LCA reconciliation M is used for inferring gene duplications, the duplication condition used is (D-i). This is correct for the LCA reconciliation between a gene tree and a species tree. However, this stringent condition is no longer appropriate as the definition of duplication events for arbitrary reconciliations. For example, consider the reconciliation f between the gene tree G and the species tree S as in Figure 1. If the original definition is used, as proposed in [8], only one duplication is inferred, which is associated with r. However, one duplication cannot produce such a gene family having the gene tree G. On the other hand, if our proposed definition is used, two duplications are inferred, one associated with r and the other with b; the implied duplication scenario is given in Figure 2.

Now, for an internal node u, we let $\delta_f(u) = 1$ if there is a duplication event associated with it, and $\delta_f(u) = 0$ otherwise. Then the *gene duplication* cost gd (f) of f is defined as:

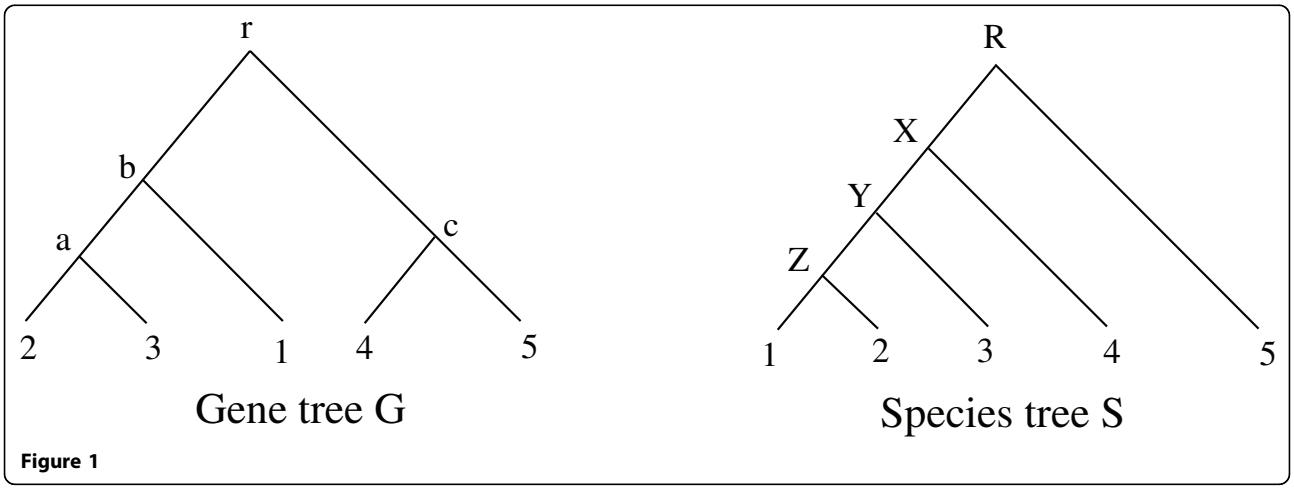

Figure 1

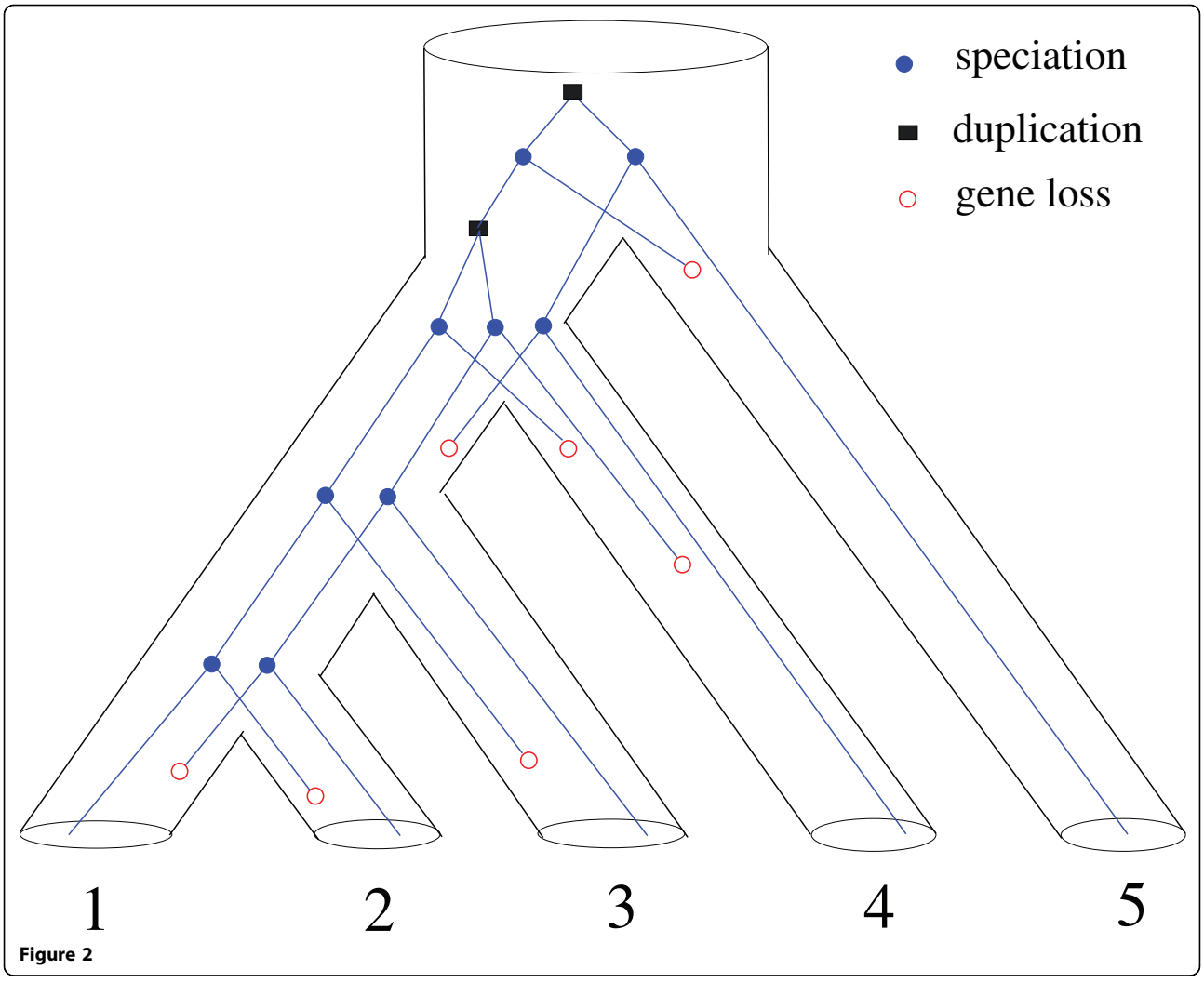

Figure 2

$$\mathrm{gd}(f) := \sum_{u \in V(G)} \delta_f(u). \quad (1)$$

Gene loss cost

Let G be a gene tree, S a species tree and f a reconciliation between G and S. Then the *number of losses* $l_f(u)$ associated to an internal vertex u is defined as:

$$l_f(u) := \begin{cases} d(f(u), f(u_1)) + d(f(u), f(u_2)), & \text{if } \delta_f(u) = 1, \\ d(f(u), f(u_1)) + d(f(u), f(u_2)) - 2 & \text{otherwise.} \end{cases}$$

Note that our definition of $l_f(u)$ is a generalization of the one introduced by Ma et al. in [6], and is consistent with the one in [20]. When f is the LCA reconciliation, our definition agrees with the traditional one [5,6]. For later use, it is often convenient to combine the two formulae in the above definition, i.e., we have:

$$l_f(u) = d(f(u), f(u_1)) + d(f(u), f(u_2)) + 2 \cdot (\delta_f(u) - 1). \quad (2)$$

For simplicity, we also set $l_f(x) = 0$ for any leaf x of G. The *gene loss* cost $\mathrm{gl}(f)$ of f is defined as:

$$\mathrm{gl}(f) := \sum_{u \in V(G)} l_f(u). \quad (3)$$

For example, for the reconciliation f in Figure 1, we have $\mathrm{gl}(f) = 7$ by noting that:

$$l_f(a) = l_f(c) = l_f(r) = 1 \text{ and } l_f(b) = 4,$$

which can also been observed from Figure 2.

Deep coalescence cost

If the discord of a gene tree G and a species tree S is due to incomplete lineage sorting, a reconciliation f

between them is measured by the deep coalescence cost [3]. Given a branch e in S, we say that there are k ($k > 0$) *extra lineages* (with respect to f) failing to coalesce on e, denoted by $\tau_f(e) = k$, if there exist $k + 1$ distinct edges (u_i, v_i) $(1 \le i \le k + 1)$ in G such that e is on the path $P(f(u_i), f(v_i))$ for each i; otherwise, we let $\tau_f(e) = 0$. The *deep coalescence* cost $\mathrm{dc}(f)$ of f is then defined as:

$$\mathrm{dc}(f) := \sum_{e \in E(S)} \tau_f(e)$$

i.e., the total number of the extra lineages with respect to f on all branches of S. For example, for the reconciliation f in Figure 1, we have $\mathrm{dc}(f) = 3$ by noting that:

$$\tau_f((R, X)) = \tau_f((X, Y)) = \tau_f((Y, Z)) = 1.$$

Results

The monotonicity of the reconciliation costs

We first have the following useful observations on the gene duplication cost.

Lemma 1 *Let f be a reconciliation between a gene tree G and a species tree S. If u is an internal vertex in G with children u_1 and u_2, then the following observations hold.*

(i): $\delta_f(u) = 1$ if and only if $f(u) \in \{f(u_1), f(u_2)\}$ or $lca(f(u_1), f(u_2)) < f(u)$.

(ii): $\delta_f(u) = 0$ if and only if $f(u_1) \ne f(u) \ne f(u_2)$ and $lca(f(u_1), f(u_2)) = f(u)$.

(iii): If $L(f(u_1)) \cap L(f(u_2)) \ne \emptyset$, then $\delta_f(u) = 1$.

(iv): If $f(u) > M(u)$, then $\delta_f(u) = 1$.

(v): If $\delta_f(u) = 0$, then $f(u) = M(u)$ and $L(f(u_1)) \cap L(f(u_2)) = \emptyset$.

Proof: Since $\delta_f(u)$ is either 0 or 1, (ii) clearly follows from (i), and (v) follows from (iii) and (iv).

To establish (i), it suffices to show that $\mathrm{lca}(f(u_1), f(u_2)) < f(u)$ if and only if $P(f(u), f(u_1))$ and $P(f(u), f(u_2))$ share a common edge. Indeed, if we have $\mathrm{lca}(f(u_1), f(u_2)) < f(u)$, then $P(f(u), f(u_1))$ and $P(f(u), f(u_2))$ share the edge that is incident to $lca(f(u_1), f(u_2))$ and its parent. On the other hand, if $P(f(u), f(u_1))$ and $P(f(u), f(u_2))$ share a common edge (s, s') with $s' < s$, then s' is a common ancestor of $f(u_1)$ and $f(u_2)$ such that $s' < s \le f(u)$. Therefore we have $L(f(u_1), f(u_2)) \le s' < f(u)$, as required.

Now we proceed to prove (iii). If $L(f(u_1)) \cap L(f(u_2)) \ne \emptyset$, then we have either $f(u_1) \le f(u_2)$ or $f(u_2) \le f(u_1)$. By symmetry, we may assume $f(u_1) \le f(u_2)$, and hence $\mathrm{lca}(f(u_1), f(u_2)) = f(u_2)$ holds. Now there are two cases to be considered, i.e., $f(u_2) = f(u)$ and $f(u_2) < f(u)$. By (i), we can conclude $\delta_f(u) = 1$ in both of them.

It remains to show (iv). Note first that we can assume $\mathrm{lca}(f(u_1), f(u_2)) > M(u)$, because otherwise we have $\mathrm{lca}(f(u_1), f(u_2)) = M(u) < f(u)$, and hence $\delta_f(u) = 1$ by (i). It follows that $f(u_i) > M(u)$ for some $i = 1, 2$. Therefore, by switching u_1 and u_2 if necessary, we can further assume $f(u_1) > M(u)$. Now we need to consider two cases: $f(u_2) > M(u)$ and $f(u_2) \le M(u)$. If $f(u_2) > M(u)$, then $f(u_1)$ and $f(u_2)$ are both contained in the path $P(f(u), M(u))$, and thus $L(f(u_1)) \cap L(f(u_2)) \ne \emptyset$ holds. On the other hand, $f(u_2) \le M(u)$ implies $f(u_2) \le f(u_1)$, and hence also $L(f(u_1)) \cap L(f(u_2)) \ne \emptyset$. Since in both cases we have $L(f(u_1)) \cap L(f(u_2)) = \emptyset$, by (iii) we obtain $\delta_f(u) \ne 1$, as required. Q.E.D

Note that (i) in the above lemma provides an additional characterization of gene duplication events. This characterization is easier for calculation while the original definition is more natural, from an evolutionary point of view. By (v) in the above lemma, if a speciation event happens at u, i.e., $\delta_f(u) = 0$, then we have $f(u) = M(u)$. This agrees with the definition of reconciliation in [20]. Now we have the following main result.

Theorem 2 *Let f and f' be two distinct reconciliations between a gene tree G and a species tree S with $f' \preccurlyeq f$; then we have:*

$$\mathrm{gd}(f') \le \mathrm{gd}(f),\ \mathrm{gl}(f') < \mathrm{gl}(f),\ \textit{and}\ \mathrm{dc}(f') < \mathrm{dc}(f). \quad (4)$$

In addition, $\delta_{f'}(u) \le \delta_f(u)$ for each $u \in V(G)$, where the equality holds for each $u \in V(G)$ if and only if $\mathrm{gd}(f') = \mathrm{gd}(f)$.

Proof: Let $D(f', f)$ be the number of vertices v in $V(G)$ with $f(v) \ne f'(v)$; then the following observation plays an important role in our proof of the theorem.

Lemma 3 *Let f and f' be the two reconciliations as given in the theorem. Then there exists a reconciliation f^* between G and S that satisfies the following three conditions:*

$$f' \preceq f^* \preceq f,\ D(f', f^*) = D(f', f) - 1 \textit{ and } D(f^*, f) = 1.$$

Proof: To establish the above lemma, we select a minimal element $v_{\min}$ (with respect to the partial order $\le$ on $V(G)$) in the set $\{u \in V(G) : f(u) \ne f'(u)\}$, which is necessarily non-empty by the assumption $f \ne f'$. In other words, $f(v) = f'(v)$ holds for any v such that $v < v_{\min}$. Now consider the map f^* defined as:

$$f^*(v) := \begin{cases} f'(v_{\min}) & \text{if } v = v_{\min}, \\ f(v) & \text{otherwise.} \end{cases}$$

Then f^* is a reconciliation between G and S. To see this, note first that f and f' are reconciliations, and hence they are leaf-preserving. Therefore we know f^* is also leaf-preserving. Let u and v be a pair of vertices in G with $u \le v$. If $u, v \in V(G) - \{v_{\min}\}$, then $f^*(u) \le f^*(v)$ because f is order-preserving. On the other hand, if $u = v_{\min}$ and $v \ne u$, then we also have:

$$f^*(v_{\min}) = f'(v_{\min}) \le f'(v) \le f(v) = f^*(v),$$

where we use the fact that f' is order-preserving and $f' \preccurlyeq f$ in the first and second inequality, respectively.

Finally, suppose $v = v_{\min}$ and $u \neq v$. By the way that $v_{\min}$ is chosen, we have $f'(u) = f(u)$, and hence also:

$$f^*(u) = f(u) = f'(u) \le f'(v_{\min}) = f^*(v_{\min}).$$

This shows f^* is order-preserving, and hence f^* is indeed a reconciliation between G and S.

It remains to show that f^* satisfies the three conditions required in the claim. Since $f' \preccurlyeq f$, from the construction of f^* we have $f' \preccurlyeq f^* \preccurlyeq f$. Noting that $v_{\min}$ is the only vertex in $V(G)$ that is mapped to different images by f and f^*, we have $D(f^*, f) = 1$. Finally, for any v in G, $f'(v) \neq f^*(v)$ if and only if $v \neq v_{\min}$ and $f'(v) \neq f(v)$. In other words, we have $D(f', f^*) = D(f', f) - 1$, which completes the proof of Lemma 3. Q.E.D.

Now it suffices to prove the theorem for the special case $D(f, f') = 1$. Indeed, if $D(f, f') = m > 1$, then by Lemma 3, there exist $m+1$ reconciliations $f_1 := f'$, f_2, ..., $f_{m+1} := f$ so that $f_i \preccurlyeq f_{i+1}$ and $D(f_i, f_{i+1}) = 1$ for $1 \le i \le m$. Applying the theorem (in the special case mentioned above) for each pair of reconciliations fi and f_{i+1}, we have $\mathrm{gd}(f_i) \le \mathrm{gd}(f_{i+1})$ for $1 \le i \le m$, and hence $\mathrm{gd}(f') = \mathrm{gd}(f_1) \le \mathrm{gd}(f_{m+1}) = \mathrm{gd}(f)$. Similarly, we can show $\mathrm{gl}(f') < \mathrm{gl}(f)$, $\mathrm{dc}(f') < \mathrm{dc}(f)$, and $\delta_{f'}(u) \le \delta_f(u)$ for each $u \in V(G)$, among which the last one implies that $\mathrm{gd}(f) = \mathrm{gd}(f')$ if and only if $\delta_f(u) = \delta_{f'}(u)$ for each $u \in V(G)$.

Now let v be the unique vertex in G with $f(v) \neq f'(v)$. Clearly, v is an internal vertex. If v is not the root, let $v_0 := p(v)$ be its parent and v_3 be its sibling, that is, the other child of v_0. The remainder argument will be divided into three cases, according to the cost measure considered.

Duplication cost case

Noting that $f(v_i) = f'(v_i) \le f'(v) < f(v)$ for $i = 1, 2$, we have:

$$\mathrm{lca}(f(v_1), f(v_2)) \le f'(v) < f(v).$$

By (i) in Lemma 1, this shows $\delta_f(v) = 1$, and hence $\delta_f(v) \ge \delta_{f'}(v)$. If v is the root of G, then we have $\mathrm{gd}(f) - \mathrm{gd}(f') = \delta_f(v) - \delta_{f'}(v) \ge 0$, as required.

Now we assume v is not the root, and proceed to show $\delta_f(v_0) \ge \delta_{f'}(v_0)$. To begin with, we can assume $\delta_{f'}(v_0) = 1$, because otherwise the inequality trivially holds. In addition, we can further assume $f(v) < f(v_0)$ and $f(v_3) < f(v_0)$, because otherwise we have $\delta_f(v_0) = 1$, which also implies the inequality. It follows that we have:

$$\delta_{f'}(v_0) = 1,\ f'(v) \le f(v) < f(v_0) = f'(v_0) \text{ and } f'(v_3) \le f(v_3) < f(v_0) = f'(v_0).$$

By (i) in Lemma 1, this leads to $\mathrm{lca}(f'(v), f'(v_3)) < f'(v_0) = f(v_0)$. Let s be the child of $f(v_0)$ so that $\mathrm{lca}(f'(v), f'(v_3)) \le s$. Since $f'(v) \le f(v) < f(v_0)$ and $f'(v_3) = f(v_3)$, s is also a common ancestor of $f(v)$ and $f(v_3)$. Therefore we have $\mathrm{lca}(f'(v), f'(v_3)) \le s < f(v_0)$. Using (i) in Lemma 1 again, we can conclude $\delta_f(v_0) = 1$, as required.

Since v is the only vertex in G with $f(v) \neq f'(v)$, for each internal vertex $g \in V(G) - \{v, v_0\}$ and its two children g_1 and g_2, we have:

$$f(g) = f'(g),\ f(g_1) = f'(g_1) \text{ and } f(g_2) = f'(g_2).$$

By definition, this implies $\delta_f(g) = \delta_{f'}(g)$ for all $g \in V(G) - \{v, v_0\}$. Combining the above observations, we can conclude that $\delta_f(u) \ge \delta_{f'}(u)$ for each $u \in V(G)$. This leads to $\mathrm{gd}(f) \ge \mathrm{gd}(f')$, where the equality holds if and only if $\delta_f(u) = \delta_{f'}(u)$ for each $u \in V(G)$.

Gene loss case

Since $f(v_i) \le f'(v) < f(v)$ holds for $i = 1,2$, we have:

$$d(f(v), f(v_i)) = d(f(v), f'(v)) + d(f'(v), f(v_i)) \text{ for } i = 1, 2. \quad (5)$$

Together with the definition of l_f, we obtain:

$$\begin{aligned} l_f(v) - l_{f'}(v) &= d(f(v), f(v_1)) + d(f(v), f(v_2)) + 2 \cdot (\delta_f(v) - 1) \\ &\quad - d(f'(v), f(v_1)) - d(f'(v), f(v_2)) - 2 \cdot (\delta_{f'}(v) - 1) \\ &= 2d(f(v), f'(v)) + 2 \cdot (\delta_f(v) - \delta_{f'}(v)). \end{aligned}$$

Since $\delta_f(v) \ge \delta_{f'}(v)$, following the proof of the duplication cost case, and $d(f(v), f'(v)) > 0$, we can conclude that $l_f(v) - l_{f'}(v) \ge 0$. If v is the root of G, then this leads to $\mathrm{gl}(f) > \mathrm{gl}(f')$, as required.

Now we assume v is not the root of G. Then we have:

$$\begin{aligned} l_f(v_0) - l_{f'}(v_0) &= d(f(v_0), f(v)) + d(f(v_0), f(v_3)) + 2 \cdot (\delta_f(v_0) - 1) \\ &\quad - d(f(v_0), f'(v)) - d(f(v_0), f(v_3)) - 2 \cdot (\delta_{f'}(v_0) - 1) \\ &= -d(f(v), f'(v)) + 2 \cdot (\delta_f(v_0) - \delta_{f'}(v_0)), \end{aligned}$$

where we use the observation that $f'(v) < f(v) \le f(v_0)$ implies:

$$d(f(v_0), f'(v)) = d(f(v_0), f(v)) + d(f(v), f'(v)).$$

Combining these results, we have:

$$\mathrm{gl}(f) - \mathrm{gl}(f') = l_f(v_0) + l_f(v) - l_{f'}(v_0) - l_{f'}(v) = d(f(v), f'(v)) + 2 \cdot (\mathrm{gd}(f) - \mathrm{gd}(f')).$$

Since $\mathrm{gd}(f) \ge \mathrm{gd}(f')$, following the proof of the duplication cost case, and $d(f(v), f'(v)) > 0$, we obtain $\mathrm{gl}(f) > \mathrm{gl}(f')$, which completes the proof of this case.

Deep coalescence case

Let $E_f(S)$ be the set of edges e in S such that there exists an edge (u,u') in G such that e is contained in the directed path from $f(u)$ to $f(u')$. Now by counting extra

lineages in terms of the edges contained in paths that have form $P(f(u), f(u'))$ for some edge (u, u') in G, we have:

$$\mathrm{dc}(f) = -\left|E_f(S)\right| + \sum_{(u,u')\in E(G)} d(f(u), f(u')). \quad (6)$$

Since $E_f(S) = E_{f'}(S)$ and $f(u) = f'(u)$ for $u \neq v$, the above formula implies:

$$\mathrm{dc}(f) - \mathrm{dc}(f') = \sum_{i=0}^{2} d(f(v_i), f(v)) - d(f(v_i), f'(v)) = d(f(v), f'(v)) > 0,$$

if v is not the root of G. Here in the second equality we use the observation that $f(v)$ is on the directed path from $f(v_0)$ to $f'(v)$, and for $i = 1,2$, $f'(v)$ is on the directed path from $f(v)$ to $f(v_i)$. If v is the root of G, then a similar argument leads to:

$$\mathrm{dc}(f) - \mathrm{dc}(f') = 2 \cdot d(f(v), f'(v)) > 0,$$

which completes the proof. Q.E.D.

Since the LCA reconciliation is the minimal element in the space of reconciliations, the above theorem leads directly to the following result.

Corollary 4 *Among all reconciliations between a gene tree G and a species tree S, the LCA reconciliation has* (*a*) *the minimum gene duplication cost*[9], (*b*) *the unique one with the optimal gene loss cost*[15]*and the optimal deep coalescence cost.*

Note that there is a close relationship among the gene duplication, gene loss and deep coalescence costs [7]. From their relationship, one can easily obtained the fact that the LCA is the unique one with the optimal gene loss cost from that it is the unique one with the optimal deep coalescence, but the reverse is not clear.

Gd-optimal reconciliations

By Corollary 4, the LCA reconciliation is the unique optimal reconciliation for the gene loss cost, as well as the deep coalescence cost. However, the LCA reconciliation may not be the unique optimal one for the gene duplication cost (see [15]). For example, for the reconciliation f in Figure 1 and the LCA reconciliation M between the gene tree and species tree in Figure 1, we have $\mathrm{gd}(f) = \mathrm{gd}(M) = 2$. Since the reconciliations with the minimum gene duplication cost, which we shall refer to as *gd-optimal reconciliations*, may not be unique, in this section we will present a characterization of them, using the theoretical results developed above.

By Theorem 2, a reconciliation f is gd-optimal if and only if $\delta_f(u) = \delta_M(u)$ holds for each vertex u in G. Based on it, we will show that there exists a unique maximal gd-optimal reconciliation M^* so that f is gd-optimal if and only if $f \preccurlyeq M^*$ holds. The reconciliation M* between a gene tree G and a species tree S can be constructed as follows. For all $u \in V(G)$ with $\delta_M(u) = 0$, M^* maps u to $M(u)$, i.e., $M^*(u) = M(u)$. For those $u \in V(G)$ with $\delta_M(u) = 1$, we shall define $M^*(u)$ recursively. If $u = r(G)$, i.e., it is the root of G, then $M^*(u)$ is defined as $r(S)$, the root of S. Otherwise, $M^*(p(u))$ has been defined, and $M^*(u)$ is defined as:

$$M^*(u) = \begin{cases} M^*(p(u)), & \text{if } \delta_M(p(u)) = 1, \\ \text{The largest vertex } s \text{ in } S \text{ satisfying } M(u) \le s < M^*(p(u)), & \text{otherwise.} \end{cases}$$

If u is a vertex in G such that $u \neq r(G)$, then $\delta_M(p(u)) = 0$ implies $M(u) < M(p(u)) \le M^*(p(u))$, hence the mapping M^* is well defined. In addition, M^* is also a reconciliation between G and S. To see this, note that if u is a leaf in G, then we have $\delta_M(u) = 0$, which implies $M^*(u) = M(u)$ and hence M^* is leaf-preserving. On the other hand, by the construction of M^*, it is order-preserving. For example, for the gene tree and species tree in Figure 1, the reconciliation M^* is defined as:

$$M^*(a) = Y,\ M^*(b) = M^*(c) = M^*(r) = R,\ \text{and } M^*(i) = i \text{ for } 1 \le i \le 5.$$

In this example, it is not difficult to check that $\mathrm{gd}(f) = \mathrm{gd}(M^*)$ holds for all $f \preccurlyeq M^*$, which also follows directly from the following general result.

Theorem 5 *Given a gene tree G and a species tree S, a reconciliation f is gd-optimal if and only if* $M \preccurlyeq f \preccurlyeq M^*$ *holds. In particular, M* is the unique maximal gd-optimal reconciliation between G and S.*

Proof: We need only to show that $\mathrm{gd}(f) = \mathrm{gd}(M)$ for a reconciliation f if and only if $M \preccurlyeq f \preccurlyeq M^*$ holds, because this implies M^* is indeed the unique maximal gd-optimal reconciliation.

To show that $\mathrm{gd}(M) = \mathrm{gd}(f)$ holds for every reconciliation f with $M \preccurlyeq f \preccurlyeq M^*$, it suffices to prove $\mathrm{gd}(M^*) = \mathrm{gd}(M)$, because together with Theorem 2, this implies $\mathrm{gd}(f) = \mathrm{gd}(M) = \mathrm{gd}(M^*)$. To this end, we need only to show $\delta_M(u) = \delta_{M^*}(u)$ for each internal vertex u in G. Now fix an internal vertex u in G. Since $M \preccurlyeq M^*$, we have $\delta_{M^*}(u) \ge \delta_M(u)$ by Theorem 2. If $\delta_M(u) = 1$, then we have $\delta_{M^*}(u) = 1 = \delta_M(u)$. Therefore it remains to consider the case $\delta_M(u) = 0$. By (ii) in Lemma 1, $\delta_M(u) = 0$ implies $M(u_1) \neq M(u) \neq M(u_2)$. Together with the construction of M^*, we have $M^*(u_1) \neq M^*(u) \neq M^*(u_2)$. Since $M(u_i) \le M^*(u_i) \le M^*(u)$ for $i = 1, 2$, we have:

$$\mathrm{lca}(M(u_1), M(u_2)) \le \mathrm{lca}(M^*(u_1), M^*(u_2)) \le M^*(u) = M(u).$$

By the construction of M, we know $\mathrm{lca}(M(u_1), M(u_2)) = M(u)$, and hence:

$$\mathrm{lca}(M^*(u_1), M^*(u_2)) = M^*(u).$$

By (ii) in Lemma 1, this shows $\delta_{M^*}(u) = 0$, as required.

To establish the other direction, assume $\mathrm{gd}(f) = \mathrm{gd}(M)$ for a reconciliation f, and we shall show $f \preccurlyeq M^*$, i.e., $f(u) \le M^*(u)$ for each internal u in $V(G)$. To this end, fix an internal vertex u in G, and denote its two children by u_1 and u_2. If $\delta_f(u) = 0$, then by (v) in Lemma 1 we have $f(u) = M(u)$, and hence $f(u) = M^*(u)$. Therefore, it remains to prove $f(u) \le M^*(u)$ for $\delta_f(u) = 1$, which will be established by induction. The base case is u being the root of G; then $M^*(u)$ is the root of S, and $f(u) \le M^*(u)$ trivially holds. For the induction step, let $u_0 := p(u)$ be the parent of u; then the induction assumption is $f(u_0) \le M^*(u_0)$. Now if $\delta_M(u_0) = 1$, then by the definition of M^* we have:

$$f(u) \le f(u_0) \le M^*(u_0) = M^*(u).$$

Otherwise, we have $\delta_M(u_0) = 0$. Together with $M \preccurlyeq f$, $M \preccurlyeq M^*$ and $\mathrm{gd}(M) = \mathrm{gd}(f) = \mathrm{gd}(M)$, this leads to $\delta_f(u_0) = \delta_{M^*}(u_0) = 0$ by Theorem 2. In view of (v) in Lemma 1, we obtain $M(u_0) = f(u_0) = M^*(u_0)$. Since $\delta_f(u_0) = 0$, (ii) in Lemma 1 implies $f(u) \prec f(u_0)$, and hence also:

$$M(u) \le f(u) < f(u_0) = M(u_0).$$

By definition, $M^*(u)$ is the largest vertex in the set $\{s : M(u) \le s \prec M^*(u_0)\}$. Since $M^*(u_0) = M(u_0)$, we can conclude $f(u) \le M^*(u)$, which completes the proof. Q.E.D.

Enumerate nearly-optimal reconciliations

Recall that there are other reconciliations having the minimum duplication cost than the LCA reconciliation. Moreover, in a biological study, a nearly-optimal reconciliation could be the correct solution to its problem. Therefore, it is of interest to study the following problem [20]: Given a positive number ε, compute the set of *nearly-optimal* reconciliations that have the duplication cost less than or equal to $\mathrm{gd}(M) + \varepsilon$, where $\mathrm{gd}(M)$ is the minimum duplication cost a reconciliation between the gene tree and the species tree can have. Such a subset of the nearly-optimal reconciliations is denoted by $\Gamma_\varepsilon(G, S, \mathrm{gd})$, which is also a subset of $\Gamma(G, S)$, the set of all reconciliations between G and S.

In this section we will present an algorithm for enumerating $\Gamma_\varepsilon(G, S, \mathrm{gd})$. To this end, we need to introduce some additional definitions. Following [20], for a vertex $u \in V(G)$, let $\mathrm{id}(u)$ be the number of vertices that precede u according to the prefix traversal of G, where the left child u_1 of a vertex $u \in V^\circ(G)$ is visited before the right child u_2. For a reconciliation f in $\Gamma(G, S)$, and a vertex $u \in V^\circ(G)$ with $f(u) \neq r(S)$, $f[u]$ is a mapping defined as:

$$f[u](v) := \begin{cases} p(f(u)) & \text{if } v = u, \\ f(v) & \text{otherwise.} \end{cases}$$

For an internal vertex u with $u \neq r(G)$, $f[u]$ is a reconciliation if and only if $f(u) \prec f(p(u))$; for the root $r(G)$ of G, $f[r(G)]$ is a reconciliation if and only if $f(r(G)) \prec r(S)$. In both cases, we will say that the reconciliation $f[u]$ is obtained from f by applying a Nearest Mapping Change (NMC) operator on u; this operator is adapted from the one introduced in [20]. Similarly, we can define $f[u_1, \ldots, u_k]$ for a sequence of (not necessarily distinct) vertices in G. Note that for a reconciliation f in $\Gamma(G, S)$ with $f \neq M$, there exists a unique sequence $u_1, \ldots, u_k$ so that $f = M[u_1, \ldots, u_k]$ and $\mathrm{id}(u_i) \le \mathrm{id}(u_{i+1})$ for $i = 1, \ldots, k-1$; now $\mathrm{id}(f)$ is defined as $\mathrm{id}(u_k)$, where u_k is the last vertex in this sequence. For completeness, we will use the convention $\mathrm{id}(M) = 0$. Finally, for a reconciliation f in $\Gamma(G, S)$, we set:

$$K(f) := \{u \in V^\circ(G) : f[u] \text{ is in } \Gamma(G, S) \text{ and } \mathrm{id}(u) \ge \mathrm{id}(f)\},$$

where $K(f)$ will be regarded as an ordered list (with the order induced by id).

The NMC operator induces a tree structure on the set $\Gamma(G, S)$: the root is M; f' is a child of f if and only if $f' = f[u]$ for some $u \in K(f)$. This tree, whose vertex set is $\Gamma(G, S)$, will be denoted by $T(G, S)$. The idea of considering a tree structure on the space of reconciliation was introduced in [20]. Clearly, by Theorem 2, the restriction of $T(G, S)$ on $\Gamma_\varepsilon(G, S, \mathrm{gd})$ is a subtree, which will be referred to as $T_\varepsilon(G, S, \mathrm{gd})$. Now we can state our algorithm as follows, which enumerates $\Gamma_\varepsilon(G, S, \mathrm{gd})$ by a traversal of $T_\varepsilon(G, S, \mathrm{gd})$. Here $\sqcup$ stands for disjoint union.

Algorithm for enumerating nearly - optimal reconciliations
Input : A gene tree G and a species tree S with $L(G) \subseteq L(S)$, and $\varepsilon \ge 0$.
Output : The set $\Gamma_\varepsilon(G, S, \mathrm{gd})$ of nearly-optimal reconciliations.

1 : Construct the LCA reconciliation M between G and S.
2 : Set $\Gamma := \{M\}$, $\mathrm{id}(M) = 0$, $\Delta(M) := 0$, $B := \{M\}$.
3 : While $B \neq \emptyset$, set $B' := \emptyset$ and **do** {
4 : For each $f \in B$, **do** {
5 : For each node $u \in K(f)$, construct the map $f' := f[u]$, and set $\mathrm{id}(f') = \mathrm{id}(u)$.
5a : Calculate $\Delta(f') := \begin{cases} \Delta(f) + \delta_{f'}(u) - \delta_f(u), & \text{if } u = r(G); \\ \Delta(f) + \delta_{f'}(u) + \delta_{f'}(p(u)) - \delta_f(u) - \delta_f(p(u)), & \text{otherwise.} \end{cases}$
5b : If $\Delta(f') \le \varepsilon$, set $B' \leftarrow \{f'\} \sqcup B'$. } **undo**
6 : Set $\Gamma \leftarrow \Gamma \sqcup B'$ and $B \leftarrow B'$. } **undo**.
7 : Output Γ.

To see the running time of the above algorithm, note first that for a reconciliation f, $K(f)$ is a subset of $V^\circ(G)$, and for each $u \in V^\circ(G)$, whether $u \in K(f)$ or not can be determined in constant time, when $\mathrm{id}(u)$ and $\mathrm{id}(f)$ are known. In addition, if δ_M is given, then line 5a and 5b can be computed in constant time; the proof of this observation will be presented in the full version of this paper. Therefore, the above algorithm runs in time $O(|V(G)| \cdot |\Gamma_\varepsilon(G, S, \mathrm{gd})|)$, plus additional preprocessing time to compute $\mathrm{id}(u)$ and $\delta_M(u)$ for each $u \in V^\circ(G)$.

Two facts prevent us from designing better algorithm for the enumeration problems. The first one concerns the boundary set $B_\varepsilon(G, S, \text{gd})$, which consists of all reconciliation f in $\Gamma(G, S) - \Gamma_\varepsilon(G, S, \text{gd})$ such that for some $f^* \in \Gamma_\varepsilon(G, S, \text{gd})$, f is a child of f^* in $T(G, S)$. In order to enumerate $\Gamma_\varepsilon(G, S, \text{gd})$, an algorithm typically needs to visit not only the reconciliations in $\Gamma_\varepsilon(G, S, \text{gd})$, but also those in $B_\varepsilon(G, S, \text{gd})$. However, $|B_\varepsilon(G, S, \text{gd})|$ could be as large as $O(|V(G)| \cdot |\Gamma_\varepsilon(G, S, \text{gd})|)$. For instance, if G and S have the same tree structure on $n+1$ leaves, then $\Gamma_0(G, S, \text{gd}) = \{M\}$ but $|B_0(G, S, \text{gd})|$ contains $n - 1$ reconciliations. Furthermore, we have $|\Gamma_1(G, S, \text{gd})| = n$ and $|B_1(G, S, \text{gd})| = \Theta(n^2)$.

The other concern is about the set $K_\varepsilon(f) := \{u \in V(G) : f[u]$ is in $\Gamma_\varepsilon(G, S, \text{gd})$ and $\text{id}(u) \geq \text{id}(f)\}$, which is needed if we want to explore $\Gamma_\varepsilon(G, S, \text{gd})$ without visiting the boundary set $B_\varepsilon(G, S, \text{gd})$. However, some properties of these two sets, $K(f)$ and $K_\varepsilon(f)$, are different. For instance, the following property of $K(f)$ is crucial to the optimal algorithm for exploring $\Gamma(G, S)$ (see Property 5 and Proposition 4 in [20]): If u is the first vertex in $K(f)$ and $f' = f[u]$, then we have $K(f) - K(f') \subseteq \{u\}$. However, this does not hold for K_ε. To see it, considering the example mentioned in the previous paragraph, and denoting the first child of $r(G)$ by r_1, then we have $K_1(M) = V^o(G) - \{r(G)\}$ while $K_1(M[r_1]) = \varnothing$.

Since $\Gamma_0(G, S, \text{gd})$ contains the gd-optimal reconciliations, the above algorithm also provides a method for enumerating all the optimal reconciliations between a gene tree and a species tree. Since $T_\varepsilon(G, S, \text{gl})$, as well as $T_\varepsilon(G, S, \text{dc})$, is also a subtree of $T(G, S)$, we also remark that it can be modified to list nearly-optimal reconciliations with respect to the gene loss or deep coalescence cost. Due to the limited space, the details of these algorithms are omitted here and one is referred to the full version of this work appearing in our personal website. As our on-going work, the algorithms presented here will be coded in C++ and evaluated by comparing them with the existing ones on simulation data.

Conclusions

To investigate all reconciliations between a gene tree and a species tree, we have generalized the LCA reconciliation to define an arbitrary reconciliation as a vertex mapping from the gene tree to the species tree. This provides a new framework for investigating various mathematical issues of the reconciliation space. It allows us to give a unified approach to study reconciliations with each of the cost models. As applications, we show that the LCA reconciliation is the unique one having the smallest deep coalescence cost, and present a characterization of the reconciliations with the minimum gene duplication cost; we also develop efficient algorithms to enumerate nearly-optimal reconciliations with each cost models. In future, we shall incorporate other evolutionary forces behind the gene tree heterogeneity, such as horizontal gene transfer and recombination, into this framework.

Acknowledgements

The work was financially supported by the Singapore MOE grant R-146-000-134-112. We thank three anonymous referees for providing constructive comments.

This article has been published as part of *BMC Bioinformatics* Volume 12 Supplement 9, 2011: Proceedings of the Ninth Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/12?issue=S9.

Authors' contributions

Both authors carried out the research and drafted the paper.

Competing Interests

The authors declare that they have no competing interests.

Published: 5 October 2011

References

1. Metzker M: **Sequencing technologies - the next generation.** *Nature Reviews Genetics* 2010, **11**:31-46.
2. Pamilo P, Nei M: **Relationship between gene trees and species trees.** *Mol. Biol. Evol* 1988, **5**:568-583.
3. Maddison W: **Gene trees in species trees.** *Syst. Biol* 1997, **46**:523-536.
4. Goodman M, Czelusniak J, Moore G, Romero-Herrera A, Matsuda G: **Fitting the gene lineage into its species lineage: A parsimony strategy illustrated by cladograms constructed from globin sequences.** *Syst. Zool* 1979, **28**:132-168.
5. Page R: **Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas.** *Syst. Biol* 1994, **43**:58-77.
6. Ma B, Li M, Zhang L: **From gene trees to species trees.** *SIAM J. Comput* 2010, **30**:729-752.
7. Zhang L: **From gene trees to species trees II: species tree inference by minimizing deep coalescence event.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2011, accepted.
8. Bonizzoni P, Della Vedova G, Dondi R: **Reconciling a gene tree to a species tree under the duplication cost model.** *Theoretical Computer Science* 2005, **347**:36-53.
9. Gorecki P, Tiuryn J: **DLS-trees: A model of evolutionary scenario.** *Theoretical Computer Science* 2006, **359**:378-399.
10. Arvestad L, Lagergren J, Sennblad B: **The gene evolution model and computing its associated probabilities.** *J. ACM* 2009, **56(2)**:1-44.
11. Chen K, Durand D, Farach-Colton M: **Notung: A program for dating gene duplications and optimizing gene family trees.** *Journal of Computational Biology* 2000, **7**:429-447.
12. Eulenstein O, Mirkin B, Vingron M: **Duplication-based measures of difference between gene and species trees.** *Journal of Computational Biology* 1998, **5**:135-148.
13. Bansal M, Eulenstein O: **The multiple gene duplication problem revisited.** *Bioinformatics* 2008, **23**:132-138.
14. Liu L, Yu L, Kubatko L, Pearl D, Edwards S: **Coalescent methods for estimating phylogenetic trees.** *Mol. Phylogenet. Evol* 2009, **53**:320-328.
15. Chauve C, El-Mabrouk N: **New perspectives on gene family evolution: Losses in reconciliation and a link with supertrees.** In *Research in Computational Molecular Biology, LNCS 5541*. Springer Berlin / Heidelberg; Batzoglou S 2009:46-58.
16. Degnan J, Rosenberg N: **Gene tree discordance, phylogenetic inference, and the multispecies coalescent.** *Trends in Ecology and Evolution* 2009, **24**:332-340.
17. Than C, Rosenberg N: **Consistency properties of species tree inference by minimizing deep coalescences.** *Journal of Computational Biology* 2011, **18**:1-15.
18. Than C, Nakhleh L: **Species tree inference by minimizing deep coalescences.** *PLoS Computational Biology* 2009, **5(9)**:e1000501.

19. Kingman J: **Origins of the coalescent. 1974-1982.** *Genetics* 2000, **156**:1461-1463.
20. Doyon J, Chauve C, Hamel S: **Space of gene/species trees reconciliations and parsimonious models.** *Journal of Computational Biology* 2009, **16**:1399-1418.
21. Doyon J, Hamel S, Chauve C: **An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework.** *preprint* 2010.
22. Arvestad L, Berglund A, Lagergren J, Sennblad B: **Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution.** *Proceedings of the eighth annual international conference on Research in computational molecular biology, RECOMB '04* 2004, 326-335.

doi:10.1186/1471-2105-12-S9-S7
Cite this article as: Wu and Zhang: Structural properties of the reconciliation space and their applications in enumerating nearly-optimal reconciliations between a gene tree and a species tree. *BMC Bioinformatics* 2011 **12**(Suppl 9):S7.